DOCUMENT RESUME

ED 447 169 TM 032 011

AUTHOR	Marzano, Robert J.
TITLE	Analyzing Two Assumptions Underlying the Scoring of Classroom Assessments.
INSTITUTION	Mid-Continent Research for Education and Learning, Aurora, CO.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	2000-01-00
NOTE	40p.
CONTRACT	RJ96006101
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; *Scoring; *Teacher Made Tests; *Test Theory

ABSTRACT

There has been little discussion of two conventions common within classroom assessment: the convention of representing student's performance on an assessment using a single score; and the convention of using the average score to summarize a student's performance over a set of assessments. This paper attempts to demonstrate that the assumptions underlying these conventions are questionable at best. The paper also attempts to demonstrate that the use of these conventions renders classroom assessment a poor feedback device. Alternative conventions that make classroom assessments more accurate and useful feedback mechanisms are presented and discussed. Suggestions made in this paper can be summarized in brief statements: (1) classroom assessments should be given throughout the cycle of learning; (2) separate scores should be assigned to each trait addressed in a given classroom assessment; (3) a common scale should be used for all traits on all assessments; (4) summary scores for the various traits should be combined in a way that acknowledges their dependence; and (5) the summary scores for a given trait should be estimated by predicting the final score in the set of scores based on some mathematical model of learning. (Contains 5 tables, 4 exhibits, and 37 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 447 169

ANALYZING TWO ASSUMPTIONS UNDERLYING THE SCORING OF CLASSROOM ASSESSMENTS

by
Robert J. Marzano

Mid-continent Research for Education and Learning
Aurora, Colorado

January, 2000

TM032011

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

This publication is based on work sponsored wholly, or in part by the Office of Educational Research and Improvement, Department of Education, under Contract Number RJ96006101. The content of this publication does not necessarily reflect the views of OERI or any other agency of the U.S. Government.

Assessment is one of the most fundamental of classroom activities. Research by Brookhart (1993, 1994) demonstrates that assessment and assessment-related activities permeate almost all aspects of classroom life including (among others) identifying timing and sequencing of activities, determination of homework assignments, timing of assessments, timing of feedback, and the like. There has been a fair amount of research on teacher grading practices (see Stiggins, Frisbie & Griswold, 1989; Terwilliger, 1989). Much of this research has focused on the various academic and non-academic factors teachers include in grades, and the various weights given to those factors. Additionally, there have been many theory-based recommendations about the construction of items for classroom assessments and how to display and interpret scores on classroom assessments (see Airasian, 1994; Haladyna, 1994, 1997; McMillan, 1997, 2000). However, there has been little if any discussion of two conventions common within classroom assessment: (1) the convention of representing a student's performance on an assessment using a single score, and (2) the convention of using the average score to summarize a student's performance over a set of assessments. This article attempts to demonstrate that the assumptions underlying these conventions are questionable at best. Additionally, this article attempts to demonstrate that the use of these conventions renders classroom assessment a poor feedback device. Finally, alternative conventions that make classroom assessments more accurate and useful feedback mechanisms are presented and discussed.

Before addressing the two conventions of interest, I should note that my comments

assume that feedback is one of the primary purposes of classroom assessment. Indeed, a strong case can be made for this assumption. To illustrate, in their meta-analyses of over 7,800 studies, Fraser et al. (1987) found that feedback was one of the most robust of classroom practices in terms of its impact on learning. In his reporting of the findings from the Fraser et al. study, Hattie (1993) commented that "the most powerful single innovation that enhances achievement is feedback. The simplest prescriptions for improving education must be 'dollops' of feedback" (p. 9). Finally, in their meta-analysis, Bangert-Drowns, et al. (1991) found that assessment has an effect size of .74 (Cohen's d) on student learning when feedback is timed properly. Effective feedback, then, is the filter through which classroom assessment is discussed in this article. It should be noted, however, that some teachers might consider other purposes for assessment equal to or greater than feedback (for discussions see Airasian, 1994; McMillan, 1997).

SINGLE SCORES ON ASSESSMENTS

The use of a single score to represent a student's achievement on a classroom assessment has strong theoretical roots, one of the strongest of which is the assumption that competence in a given domain is governed by a single trait. Indeed, this assumption underlies most discussions of true score theory. For example, in one of the foundational discussions of true score theory, Lord (1959) notes that "A mental test is a collection of tasks; the examinee's performance on these tasks is taken as an index of his standing along some psychological dimension" (p. 473). Implied in Lord's comment is

the assumption that a single dimension or a single trait underlies a collection of tests within a given domain. Other discussions of true score theory are more explicit regarding the single trait assumption. For example, Gulliksen (1950) notes that true score theory assumes a monotonic relationship (most commonly isotonic) between scores on a test and a single underlying latent trait (p. 28). Similarly, Magnusson (1966) explains that: "The trait measured by a certain performance test can be represented by a latent continuum, an ability scale on which every individual takes up a certain position. The position an individual takes up on the ability scale determines. . . his true score on the test, his position on the true-score scale" (p. 63). Finally, the single trait assumption is evident in many IRT models. To illustrate, Hambleton, et al. (1991) note that ". . . a common assumption of IRT models is that only one ability is measured by a set of items in a test" (p. 9). Indeed, a fundamental assumption underlying an item characteristic curve (ICC) is that an increase in the level of proficiency in a unidimensional trait, increases the probability of correctly answering an item designed to measure that trait (Hambleton, et al., 1991, p. 7).

Certainly, it is possible to construct tests in such a way that performance on them is a function of a single underlying trait. One hopes that those who construct standardized tests have the time, expertise, and resources to craft items in such a way that they are unidimensional. However, this is probably not the case with assessments designed by classroom teachers. To illustrate, consider Exhibit 1 which contains an adaptation of a classroom assessment (a quiz) designed by a junior high school science teacher.

Exhibit 1
Science Quiz

The table below shows the temperature and precipitation (rain or snow) in five different towns on the same day.

	Town A	Town B	Town C	Town D	Town E
Low Temperature	13°c	-9°c	22°c	-12°c	10°c
High Temperature	25°c	-3°c	30°c	-4°c	12°c
Precipitation	0cm	5cm	2cm	0cm	10cm
Humidity	Low	High	Medium	Low	High

1. Which town had the highest temperature?
2. Which town had the most precipitation?
3. Which town or towns had a combination of high humidity and high precipitation?
4. Which towns are the most likely to be located close to each other?
5. Imagine that the table was turned on its side so that the towns (A, B, C, D, and E) were the rows and the information about temperature, precipitation, and humidity was reported in the columns. Would this make it easier or harder to read the table? Explain your answer.
6. Pick one town that probably received snow and two that probably did not but for different reasons. Explain why you think each of the three towns did or did not receive snow.
7. Explain what might have happened if the low temperature in Town E had dropped to -5°c.
8. Explain the relationship between humidity and precipitation if there is one.

As described by the teacher, this quiz was used to assess student achievement within a unit on weather. An analysis of the eight items on this quiz indicates that performance on this assessment is a function of at least two traits; items 1, 2, 3, and 5 appear to address the ability to read tables; items 4, 6, 7, and 8 appear to address an understanding of the formation of snow and its relationship to temperature. Assuming

that this is a valid grouping for these items, consider the interpretation of the overall scores for two students on the quiz. Student A answers items 1, 2, 3, and 5 correctly and student B answers items 4, 6, 7, and 8 correctly. Both receive the same overall score on the quiz (under the assumption that all items were equally weighed). However, the students would most probably have very different levels of understanding and skill relative to the two traits addressed in the assessment.

This example implies that the convention of designing assessments that measure multiple traits and then using a single score to designate achievement on an assessment can mask extreme differences in student competence. To illustrate, consider Exhibit 2 which depicts a hypothetical set of assessments that might be administered over a nine-week grading period.

Exhibit 2
Assessment and Trait Coverages Across Nine Weeks

Assessments	Trait #1	Trait #2	Trait #3	Trait #4	Trait #5
1. Homework #1 (15 points)	5 points	5 points	5 points		
2. Homework #2 (15 points)	5 points	5 points	5 points		
3. Quiz #1 (30 points)	10 points	10 points	10 points		
4. Homework #3 (20 points)	5 points	5 points	5 points	5 points	
5. In-class Assignment (40 points)	5 points	5 points	10 points	10 points	10 points
6. Performance Task (60 points)	10 points	10 points		20 points	20 points
7. Quiz #2 (30 points)		5 points	10 points		15 points
8. Homework #4 (25 points)				15 points	10 points
9. Homework #5 (25 points)		10 points		10 points	5 points
10. Final Exam (90 points)	20 points	20 points	15 points	15 points	20 points
Total 350 points	60 points	75 points	60 points	75 points	80 points

The nine assessments depicted in Exhibit 2 address five different traits. For each assessment a certain number of points applies to specific traits. To demonstrate the possible effects of these traits differentially nested within assessments, again, consider the hypothetical scores for two students across these nine assessments. This is depicted in Exhibit 3 (students A and B).

Exhibit 3
Scores for Two Students

Student A						
Assessment	**Total Scores for Student**	**Student's Scores on Trait #1**	**Student's Scores on Trait #2**	**Student's Scores on Trait #3**	**Student's Scores on Trait #4**	**Student's Scores on Trait #5**
Homework #1	10/15	5 out of 5	3 out of 5	2 out of 5		
Homework #2	12/15	4 out of 5	5 out of 5	3 out of 5		
Quiz #1	25/30	8 out of 10	10 out of 10	7 out of 10		
Homework #3	18/20	5 out of 5	5 out of 5	4 out of 5	4 out of 5	
In-class Assignment	32/40	5 out of 5	5 out of 5	5 out of 10	8 out of 10	9 out of 10
Performance Task	40/60	10 out of 10	10 out of 10		6 out of 20	14 out of 20
Quiz #2	25/30		5 out of 5	5 out of 10		15 out of 15
Homework #4	20/25				11 out of 15	9 out of 10
Homework #5	18/25		8 out of 10		6 out of 10	4 out of 5
Final Exam	85/90	20 out of 20	18 out of 20	10 out of 15	15 out of 15	20 out of 20
	285/350 (81%)	57/60 (95%)	69/75 (92%)	38/60 (63%)	50/75 (67%)	71/80 (89%)

Student B						
Assessment	**Total Scores for Student**	**Student's Scores on Trait #1**	**Student's Scores on Trait #2**	**Student's Scores on Trait #3**	**Student's Scores on Trait #4**	**Student's Scores on Trait #5**
Homework #1	6/15	0 out of 5	1 out of 5	5 out of 5		
Homework #2	8/15	3 out of 5	2 out of 5	3 out of 5		
Quiz #1	20/30	4 out of 10	8 out of 10	8 out of 10		
Homework #3	15/20	2 out of 5	4 out of 5	5 out of 5	4 out of 5	
In-class Assignment	35/40	5 out of 5	3 out of 5	10 out of 10	10 out of 10	7 out of 10
Performance Task	50/60	8 out of 10	8 out of 10		18 out of 20	16 out of 20
Quiz #2	23/30		4 out of 5	10 out of 10		9 out of 15
Homework #4	22/25				14 out of 15	8 out of 10
Homework #5	25/25		10 out of 10		10 out of 10	5 out of 5
Final Exam	81/90	18 out of 20	16 out of 20	15 out of 15	12 out of 15	20 out of 20
	285/350 (81%)	40/60 (67%)	56/75 (75%)	56/60 (93%)	68/75 (91%)	65/80 (81%)

Both students receive the same composite score of 81% or 285 out of 350 possible points. In many grading schemes, this composite score would translate to a grade of C+ or B–. Either way, one might assume that both students have attained the same level of expertise relative to the traits addressed in the unit. However, an inspection of the composite scores for the five traits provides a very different picture of the two students. Student A has obtained relatively high scores on traits 1, 2, and 5 (i.e., 95%, 92%, and 89% respectively) and relatively low scores on traits 3 and 4 (63% and 67% respectively). Student B demonstrates a very different pattern of achievement across the five traits. He has performed relatively well on traits 3 and 4 (93% and 91%) relatively poorly on

traits 1 and 2 (67% and 75%) and moderately on trait 5 (81%). The overall composite score masks the difference across traits.

Of course, the example above assumes that the traits addressed in a grading period are independent. In all likelihood, this would not be the case. That is, a reasonable assumption is that traits addressed during a grading period are correlated given that all are subcomponents of a common theme (i.e., weather). However, correlated traits do not necessarily make for similar achievement profiles from student to student relative to those traits. Additionally, correlations between traits are generally much lower than might be expected, even within domains that would logically seem to have highly related component elements. To illustrate, consider the domain of writing. Ransdell and Levy (1996) studied the relationship among five elements of writing (i.e., purpose and audience, word choice, organization, style, and mechanics) with college students. An analysis of their findings indicates that the range of correlation was .23 to .73 and the average correlation (using Fisher Z transformations) was .463. The traits within the domain of reading are also thought to be highly correlated, yet an analysis of the findings reported by Abbott and Berninger (1999) indicates that among six presumably highly dependent component skills (i.e., word attacks, phonemic analysis, comprehension, and so on), the average correlation (for fourth grade students) was .28 with a range of -.07 to .57. In short, the assumption that component skills within a given domain are highly correlated might not be accurate.

Even when highly correlated traits are involved, there can be a great deal of variation in intertrait scores. To illustrate, consider Table 1 which contains hypothetic scores on two traits where these traits are correlated at three levels (r = .7, .5, .3).

Table 1
Scores on Two Traits at Three Levels of Correlation

	Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	$\bar{x}$	SD
r = .7	Trait #1	45	67	90	96	87	75	66	84	74	78	82	86	81	78	77	77.73	12.11
	Trait #2	66	70	82	85	91	84	70	77	82	83	76	79	85	72	85	79.13	7.09
	Abs. Diff.	21	3	8	11	4	9	4	7	8	5	6	7	4	6	8	7.40	4.36
r = .5	Trait #2	68	73	80	82	93	85	72	75	84	85	74	77	86	70	87	79.40	7.30
	Abs. Diff.	23	6	10	14	6	10	6	9	10	7	8	9	5	8	10	9.40	4.41
r = .3	Trait #2	69	75	78	80	95	86	75	71	87	87	73	75	86	68	90	79.67	8.31
	Abs. Diff.	24	8	12	16	8	11	9	13	13	9	9	11	5	10	13	11.40	4.41

In constructing Table 1, the scores for trait #1 have been held constant, but the scores for trait #2 were changed to reflect the change in relationship (i.e., correlation). As one would expect, the absolute value of the differences between trait #1 and trait #2 becomes larger as the correlation decreases. In other words, the pattern of trait scores from student to student becomes increasingly diverse as the correlation between traits goes down. This means that even in the case where traits are correlated, reporting composite scores will mask the true patterns of student achievement. Using an IRT model, Bolt (1999) found that students who are in the middle of the competence distribution for two correlated traits will suffer the greatest amount of profile distortion.

This finding was consistent for correlations of .3, .5, and .7.

Application Considerations

The most straightforward recommendations relative to the issue of scoring multiple traits is to report more than one score on a multi-trait classroom assessment. To illustrate, reconsider the science quiz depicted in Exhibit 1. The teacher would record two scores for each student – one score for the skill of reading tables; the other for the topic of precipitation. Over a grading period, a teacher would keep track of students' scores on the various traits addressed. (The technical support necessary for this level of detail in record keeping is discussed in the third part of this article.) Of course, keeping track of scores on separate, yet correlated categories of knowledge and skill brings up the issue of combining the scores at the end of a grading period. There are at least two approaches to this issue.

One approach is to combine trait scores in a way that reflects their intercorrelations. Anastasi and Urbina (1997) recommend the use of a linear multiple regression equation. To illustrate, assume it is known that the equation representing the relationship between three traits addressed in a mathematics class is the following:

$$\text{Overall Math Achievement} = .21\,(\text{Trait \#1}) + .21\,(\text{Trait \#2}) + .32\,(\text{Trait \#3}) + 26$$

Also assume that a given student has received the following scores on the three traits: trait #1 = 75, trait #2 = 60, and trait #3 = 90. The overall mathematics achievement score for the student would be:

$$.21\ (75) + .21\ (60) + .32\ (90) + 26 = 83.15$$

An issue not addressed by Anastasi and Urbina is the manner in which the regression weights and the constant are to be computed. (Anastasi and Urbina discuss the use of a multiple regression equation in the context of achievement batteries. In such cases, the regression weights and the constant would be computed using data from the norming sample.) Central to this issue is the fact that the regression weights and constant are parameter estimates. The issue, then, can be framed as whether the parameter estimates pertain to a class of students considered as a group, or whether parameter estimates pertain to individual students.

At a surface level, one might reason that the more data points on which parameter estimates are based, the more precise the estimates will be. This would suggest that a single regression equation should be computed combining the data sets for all students in a class. However, this approach assumes that the correlation matrix for the traits is fixed, and the variation from student to student is due to sampling errors associated with occasions, tasks, or both. Stated differently, this approach assumes that the relationship among traits is the same from student to student. A contrary position

would be that the degree of relationships between traits is unique from student to student. This position would argue for separate regression equations for each student. It is the latter position that seems most probable from a cognitive perspective.

Anderson (1995) reviews a great deal of the research on the strength association between elements of knowledge and skill. Virtually all of it (or at least the vast majority of it) supports the position that the true correlation between traits within a given domain is different from student to student. To illustrate using one example, consider the Rescorla-Wagner theory. Anderson (1995) explains that in 1972 psychologists Rescorla and Wagner formulated a rule for associative learning that can be stated as follows:

$$\Delta V = a\,(b - V)$$

In this equation, ΔV stands for the change in strength of associations between two elements – in the context of the current discussion, two traits within a given domain. V refers to the current strength of association, b refers to the maximum strength of association, and a refers to the rate at which an association is made. Stated in sentence form, the Rescorla–Wagner equation says that the change in the strength of association between two traits is a function of the rate at which an individual forms an association between the two, multiplied by the difference between the maximum strength of association between the two and the individual's current strength of association.

From this equation, it follows that for two students to have the same strengths of association between two traits, they would have to (1) have the same rate, a, of making an association between the two traits, and (2) start their learning in a given class with the same initial strengths of association, V. (One can assume that the maximum strength of association, b, would be the same for all students given that it depends on the nature of the content more than the nature of the learner.) The co-occurrence of these two events is highly unlikely. In fact, it is more probable that the parameters a and b in the Rescorla-Wagner equation are normally distributed in a given class of students which would imply that the distribution of strengths of association at any one point in time between two traits would be normally distributed. In short, it is probably unreasonable to assume that the correlations between traits are the same from student to student and discrepancies between trait correlations from student to student are functions of error. Thus, computing a multiple regression equation for each student seems advisable.

A second approach to summarizing performance across traits is to use a form of profile analysis. Anastasi and Urbina note that this approach "involves the establishment of a minimum cutoff score" for each trait (p. 159). In the present context, then, a profile analysis approach would require the identification of specific "cut-scores" on each of the traits assessed throughout a grading period. The obtained scores on the traits would not be combined in any mathematical fashion; consequently, sidestepping the issue of the strength of relationships between traits. Yet the set of traits would be addressed as

a group, in that a low score (i.e., a score below the cut-score" on any one trait) would be interpreted as a lack of competence relative to the entire set. Borrowing a term from formal logic, this approach has been referred to as "conjunctive" in that the conjunction of specific events (i.e., attainment of minimum scores on all traits) is required as evidence for a specific level of achievement within the entire set (Plake, Hambleton, & Jaeger, 1995). Conversely, the multiple regression approach is more "compensatory" in that high scores on one trait can offset low scores on other traits, particularly when a student obtains a high score on a trait with a large regression weight.

It is instructive to note that Cronbach, et al. (1997) caution against the use of conjunctive rules for combining assessments across traits. Speaking from the perspective of generalizability theory, they provide the following illustration: Assume that a five-point scale with half-point intervals (0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0) is being used to assess students on a given trait. Also assume that each student has been assessed using six tasks, and each task has a standard error of .7. Finally, assume that a rule has been established that students must obtain a minimum score of 1.0 on all six tasks to "pass." Cronbach et al. demonstrate that within this scenario, a student with a universe score of 3.0 will have a 22 percent chance to obtain one score below 1.0. Although Cronbach et al.'s example applies to multiple tasks for a single trait, the same basic logic will apply to scores across multiple traits. Each score will have an associated standard error. The more traits that are judged conjunctively, the higher the probability that one of the scores will take on an extreme value. Haladyna and Ross (1999-2000) have made

similar cautions about the use of conjunctive rules.

AVERAGING

The second convention that works against the utility of classroom assessment as a viable feedback mechanism, is the use of the average score as a summary of student performance over time. There are many reasons why the convention of averaging scores is widely accepted as the best way to summarize student achievement, not the least of which is classical true score theory. Specifically, classical test theory as described by Lord & Novick (1968), Gulliksen (1950), and Magnusson (1966) is grounded in the well recognized formula:

$$X = t + e$$

where X is the observed score on a given assessment (to use terms cogent to the current discussion), t is the true score associated with that assessment and e is the error score associated with that assessment. Those with even a passing familiarity with measurement theory in education are well aware of the assumptions underlying the error component of this model; namely, that the error component is a random, latent variable that is independent of the error component of other assessments of the same trait (see Traub, 1997). Given these characteristics, it follows that summing over multiple assessments will decrease or theoretically remove the error score component. As Magnusson (1966) notes: "The greater number of parallel tests we administer, the

greater the chances are that the random errors will cancel each other out. The sum of the error scores will be zero for an infinite number of parallel tests" (p. 64). Of course, this reasoning supports the use of the average as a summary score for a set of assessments on a given trait – the more scores that are averaged, the higher the probability that the error component cancels out.

The assumptions underlying the true score component of the model are also well known by educators although there are variations in descriptions of a true score. Lord (1959) explains that a true score is "frequently defined as the average of the scores that the examinee would make on all possible parallel tests if he did not change during the testing process" (p. 473). Magnusson (1966) describes true score in the following way: ". . . the true score which can be predicted with complete certainty from the latent continuum is the same for every individual from one parallel test to the other" (p. 63). Gulliksen (1950) defines true score for a given student as ". . . the limit that the average of his scores on a number of tests approaches, as the number of parallel tests. . . increases without limit." (p. 28)

Common to most definitions is that the true score for a given individual on a given trait is constant from assessment to assessment. Again, this assumption provides a strong case for averaging. However, this assumption is commonly violated within classroom assessment and, consequently, renders averaging an imprecise summary statistic. This point is illuminated by a consideration of the concepts of formative and summative

assessments.

The distinction between formative and summative assessment was first popularized by Scriven in 1967 as part of an AERA monograph series on evaluation. Scriven's original message was that a distinction should be made between programs that are being formulated versus programs that have evolved to their final state. Evaluation takes on different characteristics and is interpreted differently in formative versus summative situations. This distinction was soon applied to the assessment of students. Specifically, formative assessment was defined as occurring while a trait is being learned. Summative assessment occurs at the end of a learning cycle (see McMillan, 2000).

One can correctly conclude that if assessments on a given trait are all summative, then a given student's true score on a given trait remains the same from assessment to assessment. In this case, the average score for the set of assessments is an unbiased estimate of the student's true score assuming that error components are uncorrelated from assessment to assessment. However, in practical terms, it is most probably true that teachers rarely have more than one summative assessment for a given trait. Typically, teachers spend a number of weeks introducing, providing practice in, and fine tuning a given trait, all along the way using formative assessments as a form of feedback to students relative to their progress. At the end of the instructional time devoted to the trait, a comprehensive assessment is administered. Following the

assumptions underlying true score theory, this single assessment would be the only measure that would be an appropriate estimate of the student's true score at the end of the instruction. This scenario creates the unfortunate situation in which a single assessment must be used to make judgments about a given student's proficiency in a given trait. The average of the formative and summative assessments would certainly not be an unbiased estimate of the student's true score because it would change from assessment to assessment assuming that learning occurs.

Another way of looking at the problems associated with the average score as a summary measure of achievement is that averaging assumes that all deviations from the mean (in the set of scores being averaged) are random, independent errors since all scores on all assessments are estimates of the same true score. Nunnally (1967) makes this assumption explicit in his discussion of obtained scores: "Obtained scores are biased estimates of true scores. Scores above the mean are biased upward. Scores below the mean are biased downward" (p. 200). Cohen and Cohen (1975) further explain that if all assessments measure the same true score, then there should be no pattern in the residuals (i.e., deviations) from the mean. However, if there is a correlation between the residuals, by definition, they are not independent and their average is not an unbiased estimate of the true score. Quite obviously, when learning is occurring, the true score for a given student is changing (increasing). Residuals from the mean will, therefore, be correlated. The average score, then, for a set of assessments that are not all summative is not an effective mathematical model for true score estimation.

Fortunately, there are ways to model true score development as learning occurs over a period of time and, consequently, utilize the information provided by both formative and summative assessments. Many psychologists have noted that most learning can be modeled using a curvilinear function, the most useful of which appears to be a power function (i.e., $y=ax^b$, where x is the predictor variable and a and b are constants). Initially described by Newell and Rosenbloom (1981), the "power law" appears to be ubiquitous, applying to a great variety of learning situations. As Anderson (1995) explains, "Since its identification by Newell and Rosenbloom, the power law has attracted a great deal of attention in psychology, and researchers have tried to understand why learning should take the same form in all experiments" (p. 196).

To illustrate, consider the data reported in Table 2 which is an adaptation of that reported by Anderson (1995). The data represents a given student's scores over a period of seven practice sessions.

Table 2
Observed and Predicted Scores Over Seven Practice Sessions

Session	1	2	3	4	5	6	7	$\bar{x}$
Observed Score	53	67	78	88	92	95	94	.81
Predicted	54.20	67.39	76.55	83.79	89.88	95.18	99.81	

As depicted in Table 2, the student began with a score of 53 and ended with a score of 94. Each of the seven observed scores contains true score and error score components. The most useful mathematical model would be the one that most accurately estimates the student's true score at the end of the seven assessments. The average for these seven scores is .81. However, as described above, the average score does not account for an increase in true score over the practice interval. However, if the student's learning does adhere to the power law, then the appropriate mathematical model to represent the student's learning at the end of the interval is a power function using the final score in the set predicted by the model as the estimate of the student's true score and the occasions (i.e., practice sessions) as the predictor variable. To demonstrate, a power function was approximated by transforming both the practice session number and the observed scores in Table 2 to their natural logs, regressing the observed scores on the session numbers and then transforming the predicted scores back to their original metric. Although this technique is an approximation only to a power function (see Motulsky, 1996), the reason for its use will be discussed in the final section of this article.

The predicted scores obtained from this procedure are presented in the third row of Table 2. The final predicted score is 95.18. If, in fact, the student's learning followed a power function, this is a viable estimate of the student's true score at the end of the learning period. It is interesting to note that the final predicted score varies greatly from the mean score (81). Even intuitively, the predicted final score appears to be a

better candidate for the final true score than is the average score.

It should be noted that the power function is not the only possible model of true score development. In fact, there is a set of potential candidates that include (1) a linear function ($y = ax + b$), (2) an exponential function ($y = ae^{bx}$), (3) a logarithmic function ($y = \log_e X$), (4) a quadratic function ($y = ax + b\,x^2 + c$), and (5) a "pseudo-power function" (mentioned previously), derived by transforming each variable into its natural log, computing a linear function ($\log_e y = a \log_e x + b$) and then transforming the predicted scores back to their original metric (y predicted = e^x).

To examine the viability of these functions, consider Table 3.

Table 3
Functions for Predicting True Scores Across Practice Sessions

Original Score	53	67	78	88	92	95	94
Power	54.20	67.39	76.55	83.79	89.88	95.18	99.91
Linear	60.32	67.22	74.11	81.00	87.89	94.79	101.68
Exponential	60.26	66.08	72.47	79.46	87.14	95.56	104.78
Logarithmic	52.96	68.92	78.25	84.88	90.02	94.21	97.76
Quadratic	52.76	67.22	78.65	87.05	92.43	94.79	94.12
"Pseudo-Power"	54.22	67.40	76.55	83.79	89.87	95.17	99.89

The first row of Table 3 contains the original set of observed scores. The second row contains the predicted scores obtained by applying a best fitting power function

through the original scores and then computing the predicted scores. The remaining rows contain the predicted scores produced from their respective functions. Inspection of Table 3 indicates that the predicted scores certainly vary to greater and lesser degrees. Again, it is interesting to note that in all cases, the final predicted score is considerably higher than the average score for the original set (81) indicating that the average score is probably always an underestimate of the true score when learning occurs across a set of assessments.

One way to judge the effectiveness of the various functions that might be used to model true score development is to examine their squared residuals from the original set. Table 4 presents the squared residuals for each function.

Table 4
Residuals for Various Functions

	Squared Residuals	Percent of Explained Variance
Power	160.80	96.01
Linear	193.68	87.29
Exponential	297.32	80.49
Logarithmic	32.21	97.89
Quadratic	1.67	99.89
Pseudo-Power	193.68	96.02

Table 4 also presents the percent of variance explained by the various regression functions. As Table 4 indicates, the power function, logarithmic function, quadratic

function, and pseudo-power function all explain more than 95 percent of the variance in the observed scores. The quadratic function explains the most variance.

In summary, then, using the average score as the summary score for a set of assessments assumes a mathematical model in which each assessment is assumed to be an estimate of the same true score. If these assessments are mostly formative, as is the case in most classroom assessment situations, this model is not appropriate. Rather, the appropriate model must depict a change in true score value due to learning. One approach to modeling true score development is to assume that learning follows a power function. Another might be to test the viability of a number of functions using the criterion of minimizing the squared residuals to select the best function. Given that an appropriate function is found or selected, the best summary score for the assessments is the final score in the set predicted by the identified function.

Application Considerations

To operationalize the modeling of true score change described above, two issues must be addressed: (1) the issue of the appropriate number of assessments over time, and (2) the need for a common scale across assessments. The issue of the appropriate number of assessments over time can be broken down into two aspects: the appropriate interval between assessments and the appropriate number of assessments.

Willett (1988) has addressed the issue of the appropriate number of assessments in his discussion of measuring change in learning. Quite obviously, the more assessments (i.e., data points) the better in terms of fitting a curve through the data. However, Willett exemplifies his approach to the measurement of knowledge change using four data points only (or four "waves" of data to use his terminology).

It seems reasonable that teachers could fit in at least five assessments for a given trait within a grading period especially when one considers the fact that a single test, quiz, and so on can be used to assess more than one trait (consider the quiz depicted in Exhibit 1). For illustrative purposes, I will assume that a teacher assigns five assessments per trait. The issue to address, then, is how accurately can a student's true score progression be estimated if only five data points are available? I will also assume that a power function has been selected as the model of true score development.

Most discussions of the "power law" of learning in psychological literature assume that the intervals between assessments are equal or nearly equal. This assumption would be difficult if not impossible to meet in the classroom. Teachers might not have the flexibility to give assessments in a set schedule (say every other day). Additionally, the hiatus produced by weekends adds another mitigating factor.

There are at least two ways of addressing the issue of unequal intervals between assessments. These are depicted in Table 5.

Table 5
Order- and Time-Based Accounting

Occasion	True Score ($y=22x^{.45}$)	Order-Based Assessment #	Order-Based Prediction	Time-Based Assessment #	Time-Based Prediction
1	22.000	1	24.288	1	21.999
2	30.653				
3	36.068				
4	41.053				
5	45.390				
6	42.271	2	43.508	6	49.269
7	52.810				
8	56.081				
9	59.133				
10	62.004				
11	64.721				
12	67.306	3	61.187	12	67.303
13	69.775				
14	72.141				
15	74.416				
16	76.608	4	77.935	16	76.605
17	78.727				
18	80.778				
19	82.768				
20	84.701	5	94.023	20	84.697
$\bar{x}$	59.970		60.188		59.975
SD	18.567		27.512		24.991

The first column in Table 5 represents 20 consecutive days during which students practice a given skill each day. The second column is entitled *true score* and is derived

by applying the following power function: $y=22x^{.45}$. In other words, the scores depicted in column two of Table 5 are those one can assume will be the true scores across 20 equally spaced practice sessions given that learning follows the power function above.

Column three of Table 5 represents one way a teacher might keep track of student achievement referred to as *order-based accounting*. Here, the teacher assigns five assessments over the 20-period interval and simply numbers the assessments consecutively without regard to differences in time intervals between them. It is easy to envision a classroom teacher doing this. That is, it is easy to imagine a teacher administering assessments when the curriculum allows, and then numbering these assessments consecutively without any regard for the intervals of time between assessments. The pertinent question relative to this discussion is how accurate is the estimate of a student's true score at the end of the 20-day period using this order-based accounting of assessments. To examine this issue, the true scores for these five assessments (column 2) were regressed on the order-based assignment numbers (column 3), and the predicted scores computed[1]. It is also important to note that the pseudo-power function described above was employed in this illustration. The predicted final score (i.e., predicted score for the 20th session) using this approach is 94.023 which is an overestimate of the true final score by 9.322.

---

[1] Certainly the five assessments given to students would not produce true scores for these five occasions. Each score would also contain an error component. However, one can assume that these components are normally distributed and independent. Consequently, they would not add any systematic variation to the five observed scores. For the purposes of this demonstration, we need not include an error component to each score although we can assume that the final predicted score will include such a component.

Column five of Table 5 depicts an alternate system of record keeping that might be referred to as a *time-based accounting* system. Here, the teacher assigns an identification number to each assessment that corresponds to the number of days students have had to practice or review a given trait. Thus, the second assessment that is given to students (in terms of its order) is given an assessment number of six because it occurs six days into the instruction/assessment cycle, the third assessment (in order) is given an assessment number of 12 because it occurs 12 days into the instruction/assessment cycle and so on. In this system, then, assessment numbers mirror the true point in the instruction/assessment cycle. When the true scores for these five assessments are regressed on the time-based assessment numbers (using the pseudo-power function), the predicted final score is 84.697 which underestimates the true final score by .004 only.

This simulation implies that the time-based system provides for a more precise estimation of a given student's true score than the order-based system of accounting. In fact, the implication is that the time-based system provides for a very accurate estimate of the final true score even in the situation where only one-fourth of the data points (i.e., 5 of 20 situations) are utilized. Again, it is important to note that both the order-based and time-based accounting methods are more accurate estimates of the final true score than is the average score. Specifically, the average score for assessments 1, 6, 12, 16, and 20 is 58.5722. The discrepancy between this average score and the final true score is far greater than that between the final true score (84.701) and the predicted final score under both accounting conditions.

The final issue to address relative to modeling true score development is scale selection. Obviously, a common metric must be used from assessment to assessment. The metric that is selected (e.g., a four-point scale or a hundred-point scale) is irrelevant as long as increases or decreases in that metric can be accurately mapped onto increases or decreases in competence in the trait being assessed. It is probably safe to say that most classroom teachers would simply add up points on an assessment and then divide by the total number of points to obtain a proportion or percentage score. While this technique does have the effect of translating all scores to a common metric, its reliance on points presents severe problems.

For the point system to work, items on a given assessment must measure equal or nearly equal intervals of achievement relative to the trait being addressed. If this is not the case, then simply adding items can distort student achievement. This problem, was dramatically illustrated in 1953 in an article by Sanders. Sanders' illustration was analogous to that below:

Item #1: |———|

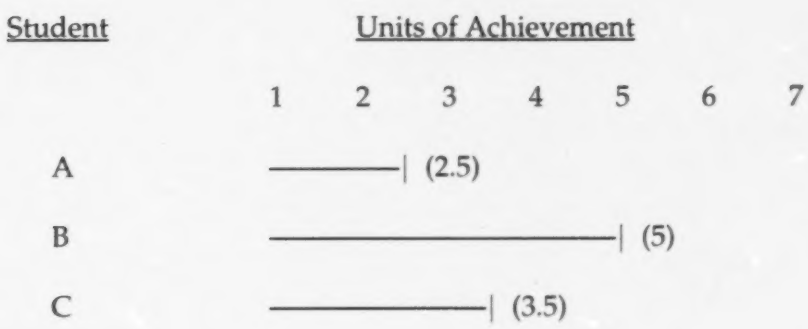
Item #2: |———|———|———|

Item #3: |———|———|

Item #4: |——|

Assume that item #1 measures a certain amount of skill relative to a given trait, item #2

measures three times the amount of skill that item #1 does; item #2 measures two times the amount of skill as item #1, and item #4 measures one-half the amount of skill as does item #1. Now assume that student A answers items #1 and #4 correctly, student B answers items #2 and #3 correctly, and student C answers items #1, #3, and #4 correctly. If one simply adds up points, it would appear that student C has the highest competence relative to the trait, and students A and B have similar competence relative to the trait. However, given the differences in the amount of achievement measured by the different items, the more accurate depiction of the achievement of the three students is presented below.



Of course, this is quite a different picture of student competence than that produced by simply adding points. Student B is actually the one who has demonstrated the most skill relative to the trait.

Prior to Sanders' demonstration of this problem in 1953, scientist S. S. Stevens had pointed out the issue in a landmark article published in 1946 entitled, *On the Theory of*

Scales of Measurement. Basically, Stevens reasoned that while it is legitimate to assume equal intervals between units when measuring physical factors such as length or width, it is erroneous to assume equal intervals between units for scales used in psychology and education. It is probably not an exaggeration to say that the stigma of "ordinal scale" dependence has plagued educational measurement ever since. As demonstrated above, the convention of adding up points and then converting total obtained points to a common metric by dividing by the total number of possible points, does little to alleviate the problems surrounding ordinal scales within educational assessment.

The problem can be framed as one of identifying a scale that represents the percentage of content attained within a given trait. Bock (1997) addresses this in his discussion of the history of item response theory. He notes that any score used to represent student performance on a given trait should reflect the percent of mastery or competence in that trait. He further argues that the percent of correct items (weighted or unweighted) in a set of items does not necessarily convey this information:

> *The concept is that in a given domain of knowledge or skills to be learned, the test should estimate the percent of competency or mastery. The estimate for a given student, called a "domain score," refers to a percentage of the domain and not to the particular collection of items on the test.* (p. 30)

The key to developing a useful scale for assessing a given trait, then, is to have some

way of mapping student responses on items in an assessment onto a description of levels of competence or mastery within a given trait. At Mid-continent Research for Education and Learning (McREL), descriptions of two general types of traits have been used quite effectively to this end (see Marzano & Kendall, 1996). Specifically, Exhibit 4 contains general descriptions of levels of understanding and skill for two types of knowledge – two types of traits.

Exhibit 4
Descriptions of Levels of Competence for Two Types of Knowledge

Declarative Knowledge	Procedural Knowledge
3 The student has a complete and detailed understanding of the information important to the trait.	3 The student can perform the skill or process important to the trait with no significant errors and with fluency. Additionally, the student understands the key features of the skill process.
2 The student has a complete understanding of the information important to the trait but not in great detail.	2 The student can perform the skill or process important to the trait without making significant errors.
1 The student has an incomplete understanding of the trait and/or misconceptions about some of the information. However, the student maintains a basic understanding of the trait.	1 The student makes some significant errors when performing the skill or process important to the trait but still accomplishes a rough approximation of the skill or process.
0 The student's understanding of the trait is so incomplete or has so many misconceptions that the student cannot be said to understand the trait.	0 The student makes so many errors in performing the skill or process important to the trait that he or she cannot actually perform the skill or process.

The left-hand portion of Exhibit 4 contains a description of levels of understanding for declarative knowledge; the right-hand side contains a description of levels of skill for procedural knowledge. Snow and Lohman (1989) note that the distinction between

declarative and procedural knowledge should be a fundamental concept in modern assessment theory. Specifically, they note that cognitive psychology has articulated many distinctions but that between declarative and procedural knowledge seems "basic for educational measurement" (p. 266). More pointedly, the distinction between declarative and procedural knowledge might help educational measurement move away from its reliance on ordinal scales in that it allows for the mapping of student responses on an assessment onto a description of development within a given trait. In Bock's words, it allows for an estimate of a student's "domain score" without the constraints imposed by points.

To illustrate, consider the scale for procedural knowledge. The bottom level of the scale certainly represents a zero point in skill – inability to execute the skill (a characteristic of ratio scales), and the top level of the scale represents mastery. The extent to which the interim two levels are equidistant between the two end points, then, determines the extent to which the scale has interval (perhaps even ratio) characteristics. At this point, no claim can be made that levels 1 and 2 on the scale are equidistant from each other and from the end points. However, a case can be made that they are better approximations to equidistant points than can be obtained using the point method. This point is best made from the perspective of their respective "frames of reference." Scales like those in Exhibit 4 start with a description of levels of understanding or skill for a given trait. The latent continuum of trait development is explicit from the outset in this scoring system. The task of the teacher scoring a student's assessment is to

accurately map the student's responses onto the explicit description of this continuum. Certainly, error will be present in this approach, but, at least, the latent continuum of the trait being measured is explicit from the outset. The point system, on the other hand, does not start with a consideration of the latent continuum for the trait. Rather, it simply assumes that the more points accumulated, the more competence has been exhibited on the trait. As illustrated above, if items measure differing amounts of knowledge or skill relative to the trait, then the total number of points might not provide an accurate mapping onto this continuum. Certainly, a great deal of research is needed to establish the conditions under which the most accurate mapping onto a trait continuum is accomplished. However, at the outset, the use of scales like those depicted in Exhibit 4, looks more promising relative to its propensity to approximate an ordinal scale than does the point method.

THE CRITICAL ROLE OF TECHNOLOGY

The suggestions made in this paper can be summarized in the following way:

1. Classroom assessments should be given throughout the cycle of instruction and learning – that is, both formative and summative assessments should be administered and considered as a set that represents learning over time.
2. Separate scores should be assigned to each trait addressed in a given classroom assessment.

3. A common scale should be used for all traits on all assessments. The use of descriptions of levels of competence for declarative and procedural knowledge was discussed.
4. Summary scores for the various traits should be combined in a way that acknowledges their dependence. Use of a multiple regression equation for each student is a viable option.
5. The summary scores for a given trait should be estimated by predicting the final score in the set of scores based on some mathematical model of learning. Power functions were discussed as appropriate models.

To implement these suggestions would be virtually impossible within current methods of record keeping in which teachers record student scores on assessments by hand. Given suggestion #2 above, the record-keeping load on a teacher would be inordinate. Similarly, suggestions 4 and 5 above would render computation impossible. However, relatively straightforward computer software can be designed or adapted to address these record-keeping and computational issues. To illustrate, consider the software designed by Strategic Learning Technologies (SLT). It allows teachers to keep track of student scores on as many as 12 traits throughout a grading period using their choice of common scales. Scores are displayed in a spreadsheet format for each student – each column representing time-ordered scores for that trait. The average score along with the predicted final score are also reported for each trait. These summary scores are

recomputed each time new trait scores are added to a student's file. The method used to compute the predicted final score is the pseudo-power function described previously. It is used because: (1) it so closely approximates a power function (as depicted in Tables 3 and 5), and (2) the regression weights and intercept are easily calculated.

While the SLT program does not implement all of the suggestions made in this article, it represents a prototype of programs to come. Ideally, in the near future, the critical process of scoring assessments and summarizing student performance will be done under the guidance of an appropriate theory base and with the use of appropriate computational algorithms. Programs like that developed by SLT that allow teachers to go beyond the use of simple weighted averages for composite scores are necessary to that end.

REFERENCES

Abbott, R. D., & Berninger, V. W. (1993). Structural equation modeling of relationships among developmental skills and writing skills in primary – and intermediate – grade writers. Journal of Educational Psychology, 85(3), 478-508.

Airasian, Peter W. (1994). Classroom assessment (2nd ed.). New York: McGraw Hill.

Anastasi, A., & Urbina, S. (1997). Psychological Testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Anderson, J. R. (1995). Learning and memory: An integrated approach. New York: John Wiley & Sons.

Bangert-Downs, R. L., Kulik, C. C., Kulick, J. A., & Morgan, M. (1991). The instructional effects of feedback in test-like events. Review of Educational Research, 61(2), 213-238.

Bock, R. D. (1997) A brief history of item response theory. Educational Measurement: Issues and Practice, 16(4), 21-33.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. Applied Measurement in Education, 12(4), 383-407.

Brookhart, S. M. (1993). Teacher's grading practices: Meaning and values. Journal of Educational Measurement, 30(2), 123-142.

Brookhart, S. M. (1994). Teachers' grading: Practices and theory. Applied Measurement in Education, 7(4), 279-301.

Cohen, J., & Cohen, P. (1975). Applied multiple regression/correlation analysis for behavioral sciences. New York: John Wiley & Sons.

Cronbach, L. J., Linn, R. L., Brennan, R. L, & Haertel, E. H. (1997). Generalizability analysis for performance assessment of student achievement or school effectiveness. Educational and Psychological Measurement, 57(3), 373-399.

Fraser, B. J., Walberg, H. J., Welch, W. W., & Hattie, J. A. (1987). Synthesis of educational productivity research. Journal of Educational Research, 11(2), 145-252.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley & Sons.

Haladyna, T., & Hess, R. (1999-2000). An evaluation of conjunctive and compensatory

standard-setting strategies for test decisions. Educational Assessment, 6(2), 129-133.

Haladyna, T. M. (1994). Developing and validating multiple-choice test items. Hillsdale, NJ: Erlbaum.

Haladyna, T. M. (1997). Writing test items to evaluate higher order thinking. Boston, MA: Allyn & Bacon.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory: Volume 2. New York: Sage Publications.

Lord, F. M. (1959). Problems in mental test theory arising from errors of measurement. Journal of the American Statistical Association, 472-479.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison & Wesley.

Magnusson, D. (1966). Test theory. Reading, MA: Addison & Wesley.

Marzano, R. J., & Kendall, J. S. (1996). A comprehensive guide to designing standards-based districts, schools, and classrooms. Alexandria, VA: Association for Supervision and Curriculum Development.

McMillan, James H. (1997). Classroom assessment: Principles and practice for effective instruction. Boston: Allyn & Bacon.

McMillan, J. H. (2000). Basic assessment concepts for teachers and administrators. Thousand Oaks, CA: Corwin Press.

Motulsky, H. (1996). The graphpad guide to nonlinear regression. San Diego: Graphpad Software Inc.

Newell, A. & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), Cognitive skills and their acquisition. Hillsdale, NJ: Erlbaum

Nunnally, J. C. (1967). Psychometric theory. New York: McGraw Hill.

Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1995). A new standard-setting method for performance assessments: The dominant profile judgment method and some field-test results. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Ransdell, S., & Levy, C. M. (1996). Working memory constraints on writing quality and fluency. In C. M. Levy & S. Ransdell (Eds.), The science of writing: Theories, methods, individual differences, and applications (pp. 93-105). Mahwah, NJ: Erlbaum.

Sanders, V. L. (1953). A comment on Burke's additive scales and statistics. Psychological Review, 60, 423-424.

Scriven, M. (1967). The methodology of evaluation. In R. F. Stake (Ed.), Curriculum evaluation: American Education Research Association Monograph Series on Evaluation, No. 1, (39-83). Chicago: Rand McNally.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 263-331). Phoenix, AZ: American Council on Education and the Onyx Press.

Stevens, S. S. (1948). On the theory of scales of measurement. Science, 103, 677-680.

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. Educational Measurement: Issues and Practices, 8(2), 5-14.

Terwilliger, J. S. (1989, Summer). Classroom standard setting and grading practices. Educational Measurement: Issues and Practice (pp. 15-19).

Traub, R. E. (1997). Classical test theory in historical perspective. Educational Measurement: Issues and Practice, 16(6), 8-14.

Willett, J. B. (1988). Questions and answers in the measurement of change. Review of Research in Education (Vol. 15, pp. 345-422). Washington, DC: American Educational Research Association.